Information Processing and Management 40 (2004) 495–514

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Automatic performance evaluation of Web search engines

Fazli Can *,[1], Rabia Nuray, Ayisigi B. Sevdik

Department of Computer Engineering, Bilkent University, Bilkent, Ankara 06800, Turkey

Received 2 January 2003; accepted 29 April 2003

Abstract

Measuring the information retrieval effectiveness of World Wide Web search engines is costly because of human relevance judgments involved. However, both for business enterprises and people it is important to know the most effective Web search engines, since such search engines help their users find higher number of relevant Web pages with less effort. Furthermore, this information can be used for several practical purposes. In this study we introduce automatic Web search engine evaluation method as an efficient and effective assessment tool of such systems. The experiments based on eight Web search engines, 25 queries, and binary user relevance judgments show that our method provides results consistent with human-based evaluations. It is shown that the observed consistencies are statistically significant. This indicates that the new method can be successfully used in the evaluation of Web search engines.

Keywords: Information retrieval; Performance; Search engine; World Wide Web

1. Introduction

The growth of the World Wide Web is an unprecedented phenomenon. Four years after the Web's birth in 1990, a million or more copies of the first well-known Web browser, Mosaic, were in use (Abbate, 1999). This growth was a result of the exponential increase of Web servers and the value and number of Web pages made accessible by these servers. In 1999 the number of Web servers was estimated at about 3 million and the number of Web pages at about 800 million (Lawrence & Giles, 1999), and three years later, in June 2002, the search engine AlltheWeb (alltheweb.com) announced that its index contained information about 2.1 billion Web pages.

* Corresponding author. Address: Department of Computer Science and Systems Analysis, Miami University, Oxford, OH 45056, USA. Tel.: +1-513-529-5950; fax: +1-513-529-1524.

E-mail addresses: canf@muohio.edu (F. Can), rabian@cs.bilkent.edu.tr (R. Nuray), ayisigi@cs.bilkent.edu.tr (A.B. Sevdik).

[1] The majority of this work has been completed when the first author was on sabbatical leave at Bilkent University.

doi:10.1016/S0306-4573(03)00040-2

There are millions of Web users and about 85% of them use search engines to locate information on the Web (Kobayashi & Takeda, 2000). It is determined that search engine use is the second most popular Internet activity next to e-mail (Jansen & Pooch, 2001). Due to high demand there are hundreds of general purpose and thousands of specialized search engines (Kobayashi & Takeda, 2000; Lawrence & Giles, 1999).

People use search engines for finding information on the Web. A Web search engine is an information retrieval system (Salton & McGill, 1983), which is used to locate the Web pages relevant to user queries (in the paper the terms page and document will be used interchangeably). A Web search engine contains indexing, storage, query processing, spider (or crawler, robot), and user interface subsystems. The indexing subsystem aims to capture the information content of Web pages by using their words. During indexing, frequent words (that, the, this, etc.), known as stop words, may be eliminated since such words usually have no information value. Various statistics about words (e.g., number of occurrences in the individual pages or in all of the indexed Web pages) are usually stored in an inverted file structure. This organization is used during query processing to rank the pages according to their relevance scores for a given query. Hyperlink structure information about Web pages is also used for page ranking (Brin & Page, 1998; Kobayashi & Takeda, 2000). The spider subsystem brings the pages to be indexed to the system. However, for Web users a search engine is nothing but its user interface that accepts queries and presents the search results. In this study our concern is text-based search engines. We also assume that users expect a selection of documents as their search result (rather than a number—e.g., freezing point of water-, a word—e.g., the name of the largest planet-, etc.). This is the case in typical Web search engine queries (Spink, Dietmar, Jansen, & Saracevic, 2001).

For assessing the performance of search engines there are various measures such as database coverage, query response time, user effort, and retrieval effectiveness. The dynamic nature of the Web also brings some more performance measure concerns regarding index freshness and availability of the Web pages as time passes (Bar-Ilan, 2002). The most common effectiveness measures are precision (ratio of retrieved relevant documents to the number of retrieved documents) and recall (ratio of retrieved relevant documents to the total number of relevant documents in the database). Measuring the search engine effectiveness is expensive due to the human labor involved in judging relevancy. (For example, one of the subjects of our experiments spent about 6 h to judge the query results and similar observations are reported in other studies (Hawking, Craswel, Bailey, & Griffiths, 2001).) Evaluation of search engines may need to be done often due to changing needs of users or the dynamic nature of search engines (e.g., their changing Web coverage and ranking technology) and therefore it needs to be efficient.

The motivation of this study is based on the fact that identifying the most effective Web search engines satisfying the current information-needs is important both at a personal and a business level. Among others, this information can be used for (a) finding more relevant documents with less effort; (b) motivating search engine providers for higher standards; (c) implementing custom-made meta-search engines. The contribution of this study is a method for automatic performance evaluation of search engines. In the paper we first introduce the method, automatic Web search engine evaluation method (AWSEEM). Then we show that it provides statistically significant consistent results compared to human-based evaluations. This shows that our method, AWSEEM, can be used instead of expensive human-based evaluations. AWSEEM uses an experimental approach in determining the effectiveness of Web search engines. For this purpose we first collect user information-needs and the associated queries, and then run these queries on several search engines. AWSEEM downloads and ranks the top 200 pages returned by each search engine for each information-need specified by the users. In the automatic method, a certain number of the most similar pages are assumed as relevant pages (pseudo-relevance judgments). It is experimentally shown that the differences in human relevance assessments do not affect the relative performance of retrieval systems (Voorhees, 2000). Based on this observation, considering these pseudo- (or automatic) relevance judgments as different human relevance assessments, the performance of search engines is evaluated. In the final stage the consistency of the automatic and human-based evaluations is measured by statistical methods.

This paper is organized as follows. In Section 2 we give an overview of related works. We then explain the experimental environment in terms of queries and search engines involved in Section 3, since this information makes the explanation of AWSEEM, which is provided in Section 4, more intuitive. Detailed experimental results and their statistical interpretation are provided in Section 5. Conclusions and future research pointers are presented in Section 6.

2. Related work

The evaluation of the text retrieval performance in static document collections is a well-known research problem in the field of information retrieval (Salton & McGill, 1983). In this study our concern is Web search engines. In the literature there are two types of search engine evaluation approaches: testimonial and shootout. Testimonials are casual studies and state the general impression obtained after executing a few queries. Shootouts are rigorous studies and follow the information retrieval measures for evaluation purposes. Our approach in this study is of the type shootout. The study reported in Jansen and Pooch (2001) provides an extensive review and analysis of current Web searching studies. It provides a rich reference list and proposes a framework for future research on Web searching.

The Gordon and Pathak (1999) study measures the performance of eight search engines using 33 information-needs. For measuring performance it calculates recall and precision at various document cut-off values (DCVs) and uses them for statistical comparisons. Intermediaries prepare the queries from the information-need descriptions of real users. The query preparation is done iteratively to achieve the best performance of individual search engines and therefore each individual search engine query used for the same information-need may be different. The user who originated the search did the assessment of the top 20 results of each search engine. The findings of the study indicate that absolute retrieval effectiveness is low and there are statistical differences in the retrieval effectiveness of search engines. The study recommends seven features to maximize the accuracy and informative content of such studies (see Table 1).

The study reported in Hawking et al. (2001) evaluates the effectiveness of 20 search engines using TREC-inspired methods with 54 queries taken from real Web search logs. The performance measures used include precision at various DCVs, mean reciprocal rank of first relevant document, and TREC-style average precision. Recall has not been used. Statistical testing reveals high intercorrelations between performance measures and significant differences between performances of search engines. Despite the time difference among the results of this study and those of Gordon and Pathak (1999) a high level of correlation is observed. This study proposes some more features

Table 1
Desirable features of Web search evaluation according to Gordon and Pathak (1999): features 1–7 and Hawking et al. (2001): features 8–11

1. The searches should be motivated by genuine information-needs of Web users
2. If a search intermediary is employed, the primary searcher's information-need should be captured as fully as and with as much context possible and transmitted to the intermediary
3. A sufficiently large number of searches must be conducted to obtain meaningful evaluations of search engine effectiveness
4. Most major search engines should be considered
5. The most effective combination of specific features of each search engine should be exploited (i.e. the queries submitted to the engines may be different)
6. The user who needs the information must make relevance judgments (Hawking et al. (2001) assumes that independent judges can do it)
7. Experiments should (a) prevent bias towards search engines (e.g., by blinding or randomizing search outputs), (b) use accepted information retrieval measures, (c) employ statistical tests to measure performance differences of search engines
8. The search topics should represent the range of information-needs over which it is desired to draw conclusions
9. Result judging should be appropriate to the type of query submitted (e.g., some queries may need a one-line answer)
10. Document presentation should be like that of a Web browser (images should be viewable, if necessary it should be possible to follow links)
11. Dead links should count as useless answers

in addition to the seven items specified in Gordon and Pathak (1999) study (they are also provided in Table 1). Note that, our approach is different than these two studies, since we aim to automate the evaluation process.

In this study, we satisfy all the features given in Table 1 except features 2 and 5. Feature 2 does not apply to us; furthermore, contrary to the suggestion of the authors, in such studies we believe that genuine user queries should be used. By this way the true everyday behavior of the search engines will be measured. Feature 5 requires expert searchers claiming that the best query formulation should be used for a specific search engine. We disagree with this, since most Web searchers in real life are casual users. So, for the experiment to be conformant with real life situations, queries should not be modified by expert searchers and should be used as they are. (Slight syntactic or stylistic modifications to a query can be allowed to make use of features specific to a search engine, such as Boolean operators.) Although we satisfy feature 6 and acknowledge the need to make relevance judgment by the original provider of the information-need, we find this to be time consuming and costly. In the Web environment, there is no standardized test collection with complete relevance judgments (Hawking et al., 2001). Being the dynamic and huge environment that it is makes it almost impossible to determine such a collection and to obtain human relevance judgments, since it would be too cumbersome for people to judge so many Web pages. Therefore in our study we suggest an automatic way to determine relevance judgments.

These studies, as mentioned above, perform evaluations using human relevance judgments. A different kind of work is the study of Mowshowitz and Kawaguchi (2002), which measures the performance of search engines using the overlap of URLs of the matching pages. This study uses the similarity between the response vector of the collection of search engines and the response vector of a particular search engine, which is defined as bias, to evaluate the performance of that search engine. The study defines the response vector of a particular search engine as the vector of URLs making up the result sets returned by that search engine and the response vector of the collection of search engines as the union of the response vectors for each of the search engines. In order to calculate bias, norm vectors for each response vector are determined by using the number of occurrences of each URL. Two ways of calculating bias are considered in this study: one by taking account of and another one by ignoring the order of URLs. The Mowshowitz and Kawaguchi (2002) study considers only the URLs retrieved by the search engines and the number of occurrences of each URL, but do not consider the content of these URLs, which we believe is necessary in the performance evaluation of search engines.

Another study is the Chowdhury and Soboroff (2002) work, which presents a method for comparing search engine performance automatically based on how they rank the known item search result. In this study, initial query–document pairs are constructed randomly. Then, for each search engine, mean reciprocal rank is computed over all query–document pairs. If query–document pairs are reasonable and unbiased, then this method could be useful. However, construction of query–document pairs requires a given directory, which may not always be possible.

There is also the Soboroff, Nicholas, and Cahan (2001) study, which involves ranking retrieval systems without relevance judgments. It proposes a new methodology in the TREC environment; however, it is related to our work because it replaces human relevance judgment with randomly selected documents from a pool. In this study, a number of documents are selected randomly to be treated as relevant. This number is taken as the average number of relevant documents appearing in the TREC pool per topic for each year. The consistency of this method with human relevance judgment is measured by experimenting on some factors (e.g., pool depth). Ranking of retrieval systems using this methodology correlates positively with official TREC rankings. It performs a relevance judgment automatically like our study; however, unlike our study, relevant document specification is done by random selection only; i.e. the contents of the documents are not considered for relevance judgments.

3. Experimental environment

3.1. User queries and information-needs

The process of measuring retrieval effectiveness requires user queries. For the construction of the queries we asked our participants—more technically known as subjects—(Fall 2001 students of CS533 students and two professors in the Computer Engineering Department of Bilkent University) to define their information-needs in English. For this purpose, they stated their needs in a form by writing a query, query subject, query description, and the related keywords. By this method we obtained 25 information-needs prepared by 19 participants (in the paper they are also referred to as users). The query evaluations were completed in the early spring of 2002.

The queries covered a broad range of topics. The topics (number of queries in that topic) are as follows: Computer science (16), education (1), Internet (2), literature (1), music (2), plants (1), sports (1), and travel (1). The average query length is 3.80 words. The average number of Boolean operators per query is 1.68. The participants only used AND and OR Boolean operators and in eight of the queries no Boolean operator is used.

3.2. Selection of search engines

Eight search engines were selected to conduct our experiment. They are AlltheWeb, AltaVista, HotBot, InfoSeek, Lycos, MSN, Netscape, and Yahoo. At first, the search engines were selected as they are in the Gordon and Pathak (1999) study. But some of the engines used in that study caused problems. OpenText, for example, could not be reached. The search engines InfoSeek and Magellan were both found to be using "overture.com", so they returned the same results. Therefore we eliminated one of them, and chose Infoseek. In the case of Excite, we observed that it is not responding to the queries submitted from our software, so it was also eliminated. After that we chose three extra search engines: AlltheWeb, MSN, and Netscape. In choosing these three search engines, we did not consider NorthernLight because with this search engine, queries, in addition to others, returned links to pages that required some access fee. We considered choosing Google because of its popularity; however, Google's search engine had been modified to prohibit outside access at the time of the experiment. This was not due to our software, but a general prohibition; for example, the software of a previous study (Mowshowitz & Kawaguchi, 2002) also was unable to retrieve any results from this search engine. However, Google's results were evaluated through Yahoo (its partner at the time of the experiments), since Yahoo presents Google's results after displaying results from its human compiled directory. In fact only the results from Google were used and not the results from the Yahoo directory. Google's results were also retrieved by Netscape, however, at the time of the experiment Netscape's results also included results from different resources. Netscape returned the results in different sections with the following order: "Most Popular Related Searches," "Partner Search Results," "Netscape Recommends," "Reviewed Web Sites," "Web Site Categories," and "Google Results". The results from first three sections were not considered as the first section displayed alternative search terms, the second section included the paid listings from Overture, and the third section presented tools and services. When no results were found in a particular section, that section did not appear. So Netscape's results cannot be used to evaluate Google, thus they are different from Yahoo's results.

3.3. Search process

In this step, we used the queries collected from our participants and submitted them to the chosen search engines through our software. It is designed to interact with the specified search engines and return their results. Some of the Boolean queries were slightly modified so that the search engines would recognize the operators. For example, the query "polygon triangulation (code or algorithm)" was submitted to AlltheWeb as "polygon triangulation (code algorithm)", since the settings of this search engine equates the words in parentheses to the Boolean OR operator.

In the case of human-based evaluation, the participants decided the relevance of the top 20 pages of each search engine in a binary fashion: they were not asked to specify any degree of relevance. To prevent possible bias, the users were just given a list of URLs obtained by taking the union of the 20 top URLs returned by all search engines. They performed the judgment task using a CGI based user interface using their Web browsers. The CGI programs lists the combined search results and saves the user relevance judgments. The participants were unable to associate the search results with search engines. When deciding on relevancy the subjects were asked to judge only the given Web page and pay no attention to the contents of the Web pages (if any) linked to the original one (i.e. not to follow links on the page).

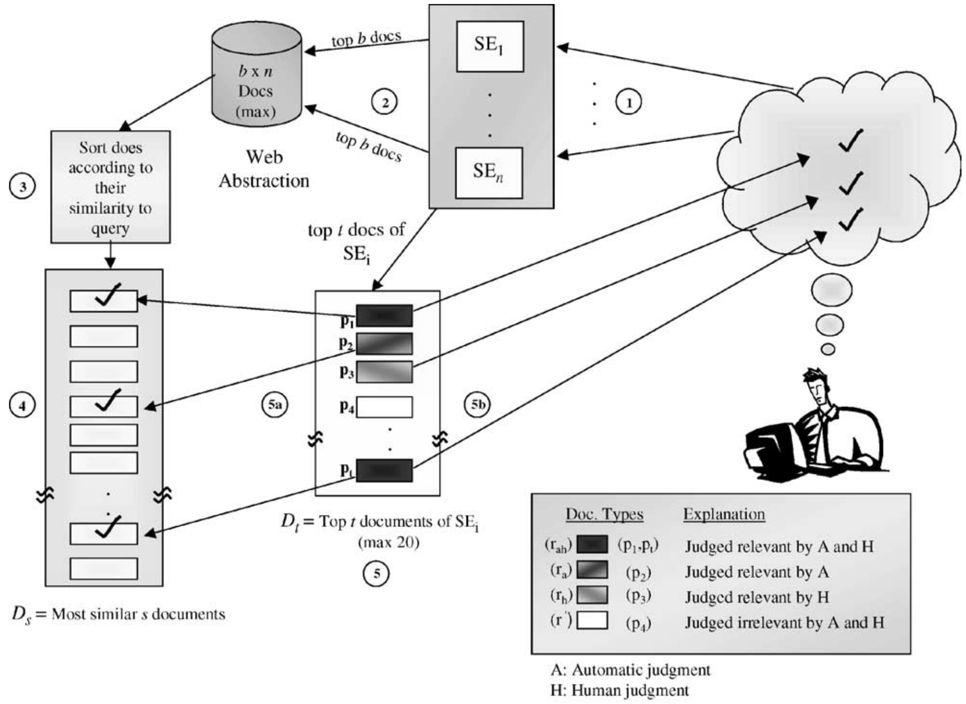
4. Automatic method AWSEEM

In AWSEEM, first all user *queries* are submitted to all search engines under consideration. After this for each query, the top b (200) Web pages of each search engine are determined. The content of these pages, if the pages are available, are downloaded and saved to build a separate document collection, or Web abstraction, for each query. In this process dead-links (i.e. unavailable pages), are regarded as useless answers. If a page is selected by more than one search engine, only one copy is downloaded and kept in the corresponding document collection. Accordingly the total number of Web pages in the collection of a query can be less than its maximum value ($n * b$). The downloaded pages are all converted to plain text (ASCII) representation. Non-word tokens, such as page structure information (e.g., HTML, pdf, ps tags), are deleted. After this, AWSEEM works like a meta-search engine and ranks the pages of a query according to their similarity (defined below) to the *user information-needs*. Note that compared to a query, *information-needs* provide a better description of what the users are looking for. In the future implementations of AWSEEM selected terms of a user designated (favorite) Web page can be used for the same purpose.

After ranking, AWSEEM processes each search engine one at a time. It examines the top t ($1 \leqslant t \leqslant 20$) pages, D_t, returned by the search engine under consideration and checks if these pages also appear in the *most similar* (the top) s number of downloaded pages (D_s). Any such document is assumed to be relevant and effectiveness measures are computed for the corresponding t value (also known as DCV). In AWSEEM, precision and recall measures are calculated at different t values and their consistencies are compared with the human-based evaluations for the same t values. Fig. 1 provides a graphical representation of the working principles of AWSEEM.

In our experiments for the value of s, 50 and 100 are used. Fifty is chosen as an approximation of the maximum number of relevant pages determined for a query in human-based evaluations (which was 44). We also use 100 to examine the effect of using a larger s value. These s values can be elaborated on in future experiments to find a more appropriate value for achieving higher consistency with human-based evaluations.

AWSEEM uses the vector space model for query–document matching (i.e., similarity calculation) and ranking (Salton & McGill, 1983). According to this model each document (downloaded Web page) is represented as a vector of terms found in the document and each query is represented as a vector of terms found in the information-need specification of the user, i.e., the query, query subject and description, and query search keywords. During this process stemming is employed and a stop word list is used to eliminate frequent words. In document ranking, similarity between a query (actually in our case the user information-need) and a document is calculated using the conventional vector product formula (Salton & McGill, 1983), which is the summation of the product of the corresponding term weights in the query and document. The weight of a term is calculated using the product of the *term frequency* (tf) and the *inverse document frequency* (idf). Term frequency is the occurrence count of a term in a document or query, whereas the inverse document frequency is a collection (in our case individual query document collection created by using the downloaded pages) dependent factor that favors terms


top b docs
b x n
Docs
(max)
Web
Abstraction
SE1
SEn
1
2
top b docs
3
Sort docs according to their similarity to query
top t docs of SEi
4
5a
5b
p1
p2
p3
p4
pt
Dt = Top t documents of SEi (max 20)
5
Ds = Most similar s documents
Doc. Types
Explanation
(rah) (p1-pt) Judged relevant by A and H
(ra) (p2) Judged relevant by A
(rh) (p3) Judged relevant by H
(r') (p4) Judged irrelevant by A and H
A: Automatic judgment
H: Human judgment

1. User submits query to n (8) search engines.
2. Top b (200) pages returned by each search engine are merged.
3. Merged pages are sorted using their similarity to the query.
4. The most similar s (50, 100) pages are assumed to be the relevant by AWSEEM.
5. Consistency of AWSEEM's relevance judgment (5a) and human judgment (5b) is checked.
   a. AWSEEM determines the relevant items in the top t pages of search engine i.
   b. User judges the top t pages of search engine i as either relevant or irrelevant.

Fig. 1. Automatic search engine evaluation process: generalized description for search engine SE_i.

concentrated in a few documents of the collection. Multiplying it with a normalization factor normalizes this value. The normalization factor is used to equalize the length of the document vectors. The study reported in Salton and Buckley (1988) suggests different choices for calculating term weights for different document and query vector sizes. We chose the query matching function defined for retrieval environments with long documents and short queries (in the Salton & Buckley (1988) study it is signified by *tfc.nfx*).

5. Experimental results and evaluation

5.1. Evaluation measures and their calculation

The effectiveness of search engines is measured in terms of precision and recall. This is done both for human judgments and AWSEEM and their results are compared to test the consistency of these two approaches. In this study, we measure "relative recall". In relative recall calculation, the denominator term, "the total number of relevant documents in the database", is replaced by "the total number of relevant documents retrieved by all search engines" either by automatic judgment or human judgment (each case, automatic and human, is handled within itself).

The precision and recall values are computed at various top *t*, or DCVs ranging between 1 and 20. (In the rest of the paper we will use the acronym DCV rather than *t* for clarity.) It is known that Web users usually look at very few result pages of search engines and therefore 20 is a suitable choice for the maximum DCV (Spink, Jansen, Wolfram, & Saracevic, 2002). Due to the small number of documents examined by users, precision gains importance over recall in search engine effectiveness evaluation. However, in some cases measuring recall can be important too. Recall shows how much of the known relevant documents the search engine managed to retrieve and if the number (or ratio) of these documents is acceptable. In our case we will only consider the first 20 documents to observe the recall performance of search engines in a typical Web search session.

As suggested by Hull (1993), the average precision or recall of each search engine can be computed at a fixed DCV using the computed precision and recall values of each query at that point. This eliminates any bias that could be caused by calculating precision or recall at only a single chosen DCV. It should be reminded that the average precision calculation used here is different than the average precision at seen relevant documents (Baeza-Yates & Ribeiro-Neto, 1999) in that we compute the average of all the precision values at each point up to the chosen DCV.

The average precision is computed in the following way. First for each search engine the average of the precisions for each query (P@*DCV*) is computed for every DCV under consideration. Then using that average precision for individual queries we are able to calculate the average precision for a search engine *at* a given DCV. The average precision *around* a DCV (*PaDCV*) is calculated by taking the average of the P@1, P@2,. . ., P@DCV values. For example, if we want to compute the average precision of a search engine *around* DCV = 3 (Pa3), we will compute the average of average precision values *at* 1, 2, 3 (i.e. Pa3 = (P@1 + P@2 + P@3)/3). An example for hypothetical search engines A, B, and C is given in Tables 2 and 3. The same procedure is also applied to compute the average recall at or *around* a DCV.

Table 2
Example showing precision values of search engines A, B, C

Query no.	Search engine	Precision at various DCVs		
		DCV = 1	DCV = 2	DCV = 3
1	A	1*	1/2	2/3*
1	B	0	1/2*	1/3
1	C	1*	1*	1*
2	A	0	0	1/3*
2	B	1*	1/2	1/3
2	C	1*	1*	2/3
3	A	0	0	0
3	B	1*	1*	2/3
3	C	1*	1/2	1/3

Precision values for three queries for $1 \leqslant \text{DCV} \leqslant 3$.

*Indicates that a relevant document is retrieved at that rank (DCV).

Table 3
P@DCV (average precision at DCV) and PaDCV (average precision around DCV) values for search engines A, B, C computed for the precision values given in Table 2 for $1 \leqslant \text{DCV} \leqslant 3$

Search engine	P@1 (P at 1)	P@2 (P at 2)	P@3 (P at 3)	Pa1 (P around 1)	Pa2 (P around 2)	Pa3 (P around 3)
A	1/3	1/6	1/3	1/3	1/4	5/18
B	2/3	2/3	4/9	2/3	2/3	16/27
C	1	5/6	2/3	1	11/12	5/6

5.2. Precision and recall values at various DCVs

Fig. 2(a) and (c) show respectively Pa10 and Ra10 results for AWSEEM(50) (i.e., when $s = 50$ in terms of Fig. 1) and the human-based evaluations. In these figures search engines are sorted according to their Pa10 or Ra10 values calculated by using human-based evaluations. Also note that the average precision and recall values for human and AWSEEM are shown on different scales to illustrate the strong association of ranks of the search engines according to their effectiveness on these two methods. Both methods display similar results in determining rank of search

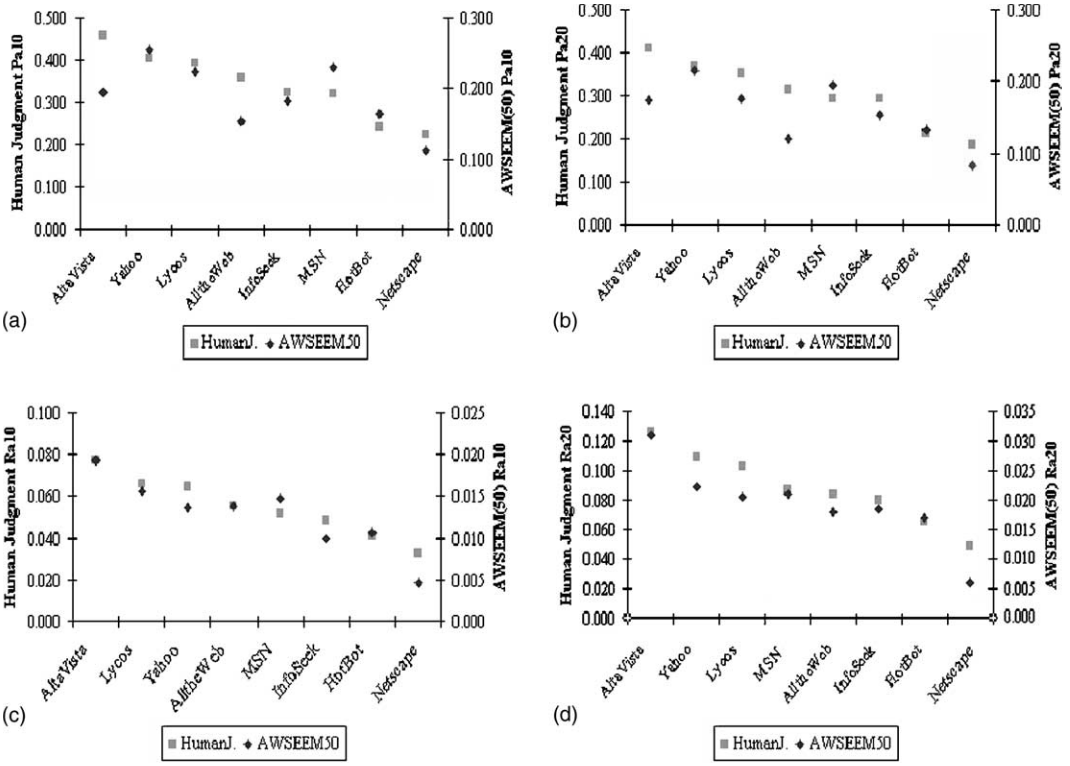


Fig. 2. (a,c) Average precision and recall around 10 for human-based evaluations and AWSEEM (50), (b,d) average precision and recall around 20 for human-based evaluations and AWSEEM (50).

engines in recall calculations, whereas the correlation in precision calculations is weaker. In the precision calculations AWSEEM ranks the first search engine differently, while the results for other search engines are very close to each other.

In Fig. 2(b) the average precision around DCV = 20 for each search engine using AWSEEM (50) is displayed with the search engines sorted by their average precision around the same DCV using human-based evaluations. In terms of human judgments, the top precision performers are AltaVista, Yahoo, and Lycos. In the case of AWSEEM (50), the best performing search engines are Yahoo, MSN, AltaVista, and Lycos (with AltaVista and Lycos having the same performance). The worst performer for both cases is Netscape. Similar results are displayed for human-based evaluation and AWSEEM in terms of ranking search engines based on precision.

If we consider the average recall calculations, human-based evaluation and AWSEEM (50) also display similar results, which can be seen in Fig. 2(d) around DCV = 20. The search engines in the figure are sorted according to their average recall values *around* DCV = 20 for human-based evaluation. Both methods show the top two performing search engines as AltaVista and Yahoo, and the worst performer as Netscape.

The calculations presented above for AWSEEM (50) have also been computed using AWSEEM (100). Fig. 3(a) and (c) show the Pa10 and Ra10 values for AWSEEM (100) and

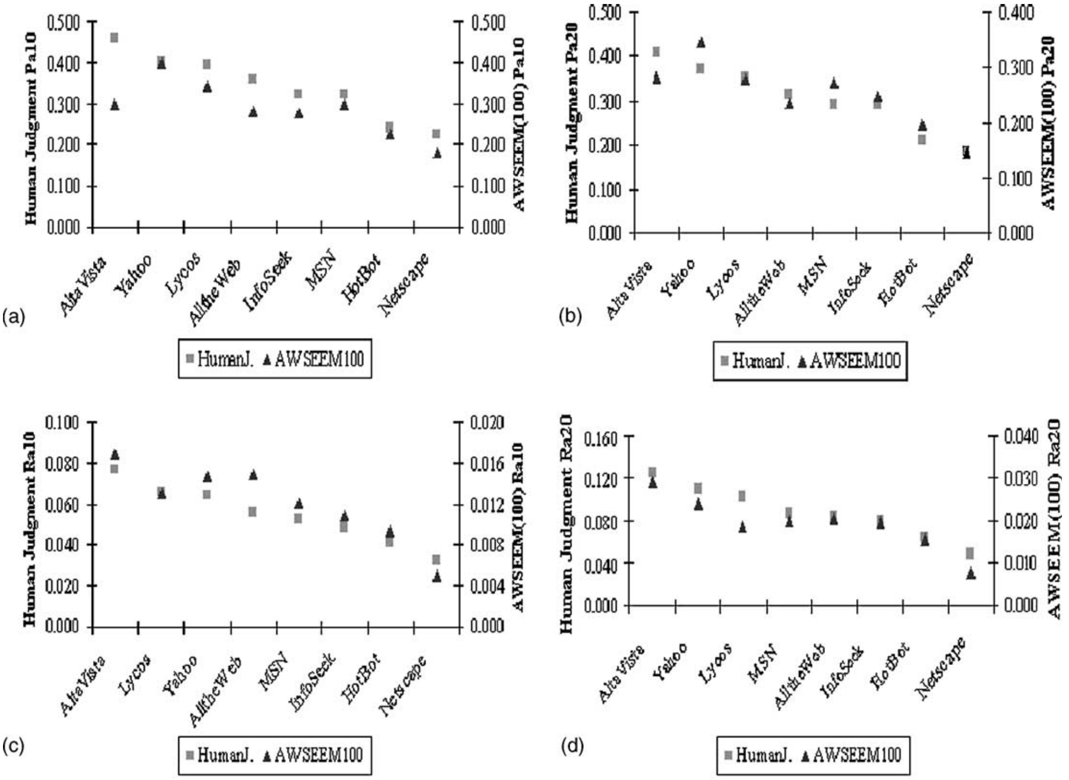


Fig. 3. (a,c) Average precision and recall around 10 for human-based evaluations and AWSEEM (100), (b,d) average precision and recall around 20 for human-based evaluations and AWSEEM (100).

human-based evaluation, where the search engines are sorted according to their Pa10 or Ra10 values calculated by using human-based evaluations. In Fig. 3(b) and (d), Pa20 and Ra20 values are displayed respectively. These figures show that both methods determine the top performing search engines as AltaVista and Yahoo, and the worst performer as Netscape. When we look at the overall results we can say that considering more pages in AWSEEM (i.e., 100 as shown in Fig. 3 vs. 50 as shown in Fig. 2) has a positive impact on the consistency of human–AWSEEM results. These figures help explain the statistical correlation between the two methods presented in the following two sections.

5.3. Statistical consistency between human and automatic judgments

Table 4 shows the number of relevant pages per query returned in the top 20 results of each search engine. This table is provided to give an insight of the actual numbers involved in the experiments. As we explained before, the relevancy of documents is separately decided by each method, so the *H* column represents the number of relevant documents in top 20 based on human

Table 4
Number of relevant pages for each query in the top 20 results of each search engine

Query	AlltheWeb		AltaVista		HotBot		InfoSeek		Lycos		MSN		Netscape		Yahoo	
–	H	A	H	A	H	A	H	A	H	A	H	A	H	A	H	A
1	11	3	2	12	2	4	3	5	1	4	0	4	1	0	0	2
2	10	4	6	3	5	3	5	3	4	0	5	2	0	2	7	4
3	3	4	6	3	3	0	6	1	6	7	5	1	9	0	7	2
4	3	5	2	2	1	4	0	6	2	3	2	6	0	2	4	9
5	5	1	7	7	8	6	8	7	9	3	8	7	1	1	9	9
6	15	4	11	4	7	3	4	2	15	5	10	2	2	0	12	4
7	1	0	9	0	7	5	0	0	3	3	0	0	2	1	2	3
8	14	1	10	2	8	2	11	2	17	2	9	2	6	2	9	4
9	15	0	12	3	13	5	16	3	15	0	16	0	13	0	15	1
10	0	11	2	8	0	0	3	8	1	5	4	10	0	1	2	6
11	4	6	5	6	1	8	1	6	7	4	1	6	0	0	6	6
12	9	1	12	2	0	0	16	4	11	1	2	1	4	0	15	2
13	7	1	3	0	0	0	6	6	4	1	5	8	6	3	5	2
14	0	0	7	3	0	0	0	0	1	0	4	0	3	1	5	1
15	2	3	5	9	4	3	3	5	1	2	4	3	2	3	4	5
16	8	1	8	2	5	3	5	4	9	1	8	2	6	3	4	5
17	2	0	3	0	0	0	4	2	0	0	10	6	7	5	9	11
18	4	6	3	5	0	0	0	6	12	11	0	6	0	0	1	5
19	4	5	6	4	4	9	3	10	7	5	4	9	0	3	12	15
20	4	6	1	4	0	0	4	8	5	11	6	9	0	0	3	8
21	0	0	5	3	5	10	0	0	5	9	6	12	4	11	6	13
22	6	13	19	12	4	7	5	5	0	0	9	12	0	1	8	7
23	7	5	12	9	5	2	3	4	9	9	5	2	4	0	7	3
24	0	0	1	7	3	2	0	0	1	7	0	0	0	3	0	0
25	0	0	10	5	0	0	12	4	0	2	0	0	0	0	12	7
Total	134	80	167	115	85	76	118	101	145	95	123	110	70	42	164	134

judgment, likewise the A column represents the results of AWSEEM (100). For example, for AlltheWeb search engine for query 1 there were 11 pages found relevant by human judgment and three pages found relevant by AWSEEM. Here, we cannot say that for performance evaluation the same set of documents is used by these two approaches; however, previous research shows that differences in human relevance judgments do not affect the relative performance of retrieval systems (Voorhees, 2000). Here, our purpose is to provide automatic (pseudo)-relevance judgments for relative performance assessment. For full details of our experimental data please refer to our Web site http://www.cs.bilkent.edu.tr/~rabian/webdata. This Web site contains various excel tables.

The consistency of AWSEEM with human judgment can be intuitively compared by using the rank of search engines according to the total number of relevant pages they retrieve for all queries (refer to the last row of Table 4). According to these values ranking of the search engines using human-based evaluation from best to worst is as follows: (1) AltaVista (with total of 167 retrieved relevant documents), (2) Yahoo (164), (3) Lycos (145), (4) AlltheWeb (134), (5) MSN (123), (6) InfoSeek (118), (7) HotBot (85), (8) Netscape (70). The ranking according to AWSEEM is as follows: (1) Yahoo (134), (2) AltaVista (115), (3) MSN (110), (4) InfoSeek (101), (5) Lycos (95), (6) AlltheWeb (80), (7) HotBot (76), (8) Netscape (42). This intuitive comparison shows that both approaches are consistent in terms of determining the best (AltaVista, Yahoo) and worst (HotBot, Netscape) performers.

To see the correlation of the results of these two methods, i.e., to test the consistency of the human and automatic results, one may use the average rank of search engines based on the precisions of the individual queries. Using this information Spearman's Rank Correlation method indicates a high level of consistency between human and AWSEEM judgments with a correlation value of 0.97, which is significant at $\alpha = 0.01$ level. We want to see the relationship of the pairs of PaDCV (RaDCV) measures for the human judgment and AWSEEM results. Spearman's correlation is suitable for variables where relationship among the ranks is of interest. Pearson r correlation is more suitable for estimating relationships between the variables themselves where the variables are reasonably normal. Since the precision values lie along a continuous scale from 0 to 1 and do not appear skewed, Pearson correlations were used to determine these relationships. Hence for the correlation of pairs of PaDCV (RaDCV) measures for various DCVs, we preferred to use Pearson r correlation.

To assess the significance of r we conduct a hypothesis test where the null hypothesis is that human and AWSEEM precision correlation is zero. We let α denote the significance level of this test; it is the error probability of rejecting the null hypothesis when the null hypothesis is correct. In the experiments high Pearson *r-values* allow us to reject the null hypothesis at lower α values. If we can reject the null hypothesis at an α of 0.05 ($r > 0.7070$), we have a strong correlation; if we can reject at α of 0.01 ($r > 0.8340$), we have a very strong correlation.

Fig. 4(a) provides Pearson r correlation between the human judgment and AWSEEM (100) results for Pa20. The scattergram shows a high degree of linear relationship between the human and AWSEEM precision results. The least squares regression line provided in the plot illustrates the variability in the data. The consistency of human-based evaluation with AWSEEM (50) and AWSEEM (100) based on precision measured by Pearson r is significant. At DCV = 20 the correlation between human judgment and AWSEEM (50) is observed as 0.7330, which is significant for $\alpha = 0.05$, and the correlation of human judgment with AWSEEM (100) is 0.8675 and

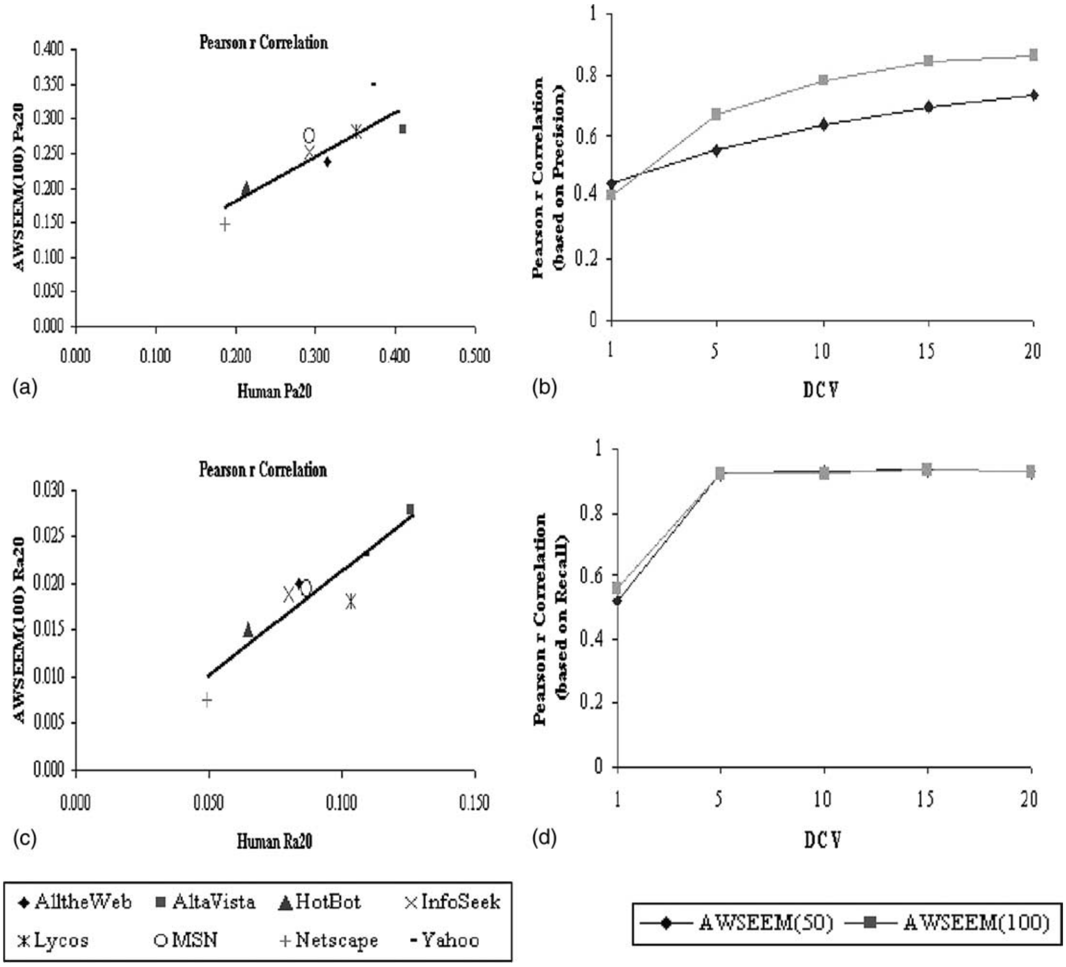


Fig. 4. (a,c) Pearson r correlation scattergram of human judgment and AWSEEM (100) for Pa20 and Ra20, (b,d) Pearson r correlation of human judgment and AWSEEM (50) and (100) for precision and recall at various DCVs.

significant for $\alpha = 0.01$. Fig. 4(b) shows the Pearson r-values for human judgment and AWSEEM for 50 and 100 based on average precision values *around* various DCVs. The figure shows that considering more pages in AWSEEM (i.e., 100 vs. 50) has a positive impact on the consistency of human–AWSEEM results. Furthermore, at DCV = 20 the highest consistency is observed. For example, AWSEEM (100) shows significant correlation ($0.7070 < r < 0.8340$) with Pa10, and more significant correlation with Pa15 and Pa20 ($r > 0.8340$). In the case of AWSEEM (50), the only significant correlation value is observed at Pa20.

If we consider average recall, the correlations of human judgment with AWSEEM (100) are strongly positive (0.9258, which is significant at $\alpha = 0.01$) as illustrated in Fig. 4(c) by using Ra20 (i.e., the average recall *around* DCV = 20) as an example. Pearson r-values for human judgment and AWSEEM for 50 and 100 based on the average recall values *around* various DCVs can be seen in Fig. 4(d). AWSEEM shows a strongly significant correlation after DCV = 5 ($r > 0.8340$).

It can be seen that the number of pages (i.e., the value of s) considered in AWSEEM does not need to be increased since the correlation of human–AWSEEM results for 50 and 100 pages is almost indistinguishable starting at DCV = 5.

5.4. Statistical distinction among search engines

Finally we performed an analysis of variance among search engines based on precision and recall scores. We calculated Tukey's highest significant differences (HSD) to determine the subsets of search engines that were statistically indistinguishable with $\alpha = 0.05$ for both methods using precision and recall values *at* DCV of 10 (i.e., P@10, R@10) and 20 (i.e., P@20, R@20). In general Tukey's HSD could not distinguish the search engines for both AWSEEM and human-based evaluations using precision and recall values *around* DCV of 10 or 20 (Pa10, Pa20, Ra10, and Ra20). Therefore, we used the precision and recall values *at* DCV of 10 and 20 (P@10, P@20, R@10, and R@20) and checked the agreement of human and AWSEEM results based on Tukey's test.

Tukey's HSD at DCV 10 using average precision values (P@10) for human-based evaluation clustered the search engines into two different subsets distinguishing between only the best performing search engine (AltaVista) and the worst performing search engine (Netscape)—see Table 5. The cluster members are shown with the letters A and B. According to Tukey's results the interpretation of a cluster in the table is that an ANOVA would find no difference among those search engines at $\alpha = 0.05$. Although Tukey's HSD for AWSEEM using the top 50 (i.e., $s = 50$) pages cannot distinguish between any of the search engines—see Table 6—using the top 100 pages it forms two clusters with Yahoo as the best performer and Netscape as the worst—see Table 7. While the best performers for human-based evaluation and AWSEEM are different, the percentage of average precision shows the same top two performers (Yahoo, AltaVista) for both methods—see Tables 5 and 7.

Tukey's HSD at DCV 20 using average precision values (P@20) for human-based evaluation clustered the search engines into two different subsets distinguishing between the best performing search engines (Yahoo, AltaVista) and the worst performing search engine (Netscape)—see Table 8. The same behavior was observed for AWSEEM using the top 50 pages—see Table 9—and

Table 5
Tukey's HSD for precision at DCV 10 P@10 ($\alpha = 0.05$) (human-based evaluation)

Search engines	N	Average precision	Subset for $\alpha = 0.05$	
			1	2
Netscape	25	0.1600	A	
HotBot	25	0.2200	A	B
MSN	25	0.2800	A	B
InfoSeek	25	0.3000	A	B
AlltheWeb	25	0.3100	A	B
Lycos	25	0.3500	A	B
Yahoo	25	0.3700	A	B
AltaVista	25	0.4100		B
Sig.			0.070	0.126

Table 6
Tukey's HSD for precision at DCV 10 P@10 ($\alpha = 0.05$) (AWSEEM (50))

Search engines	N	Average precision	Subset for $\alpha = 0.05$ 1
Netscape	25	7.200E−02	A
AlltheWeb	25	0.1000	A
HotBot	25	0.1200	A
Lycos	25	0.1500	A
InfoSeek	25	0.1500	A
AltaVista	25	0.1800	A
Yahoo	25	0.1900	A
MSN	25	0.2000	A
Sig.			0.072

Table 7
Tukey's HSD for precision at DCV 10 P@10 ($\alpha = 0.05$) (AWSEEM (100))

Search engines	N	Average precision	Subset for $\alpha = 0.05$	
			1	2
Netscape	25	0.1200	A	
HotBot	25	0.1700	A	B
AlltheWeb	25	0.2000	A	B
Lycos	25	0.2400	A	B
InfoSeek	25	0.2400	A	B
MSN	25	0.2800	A	B
AltaVista	25	0.3000	A	B
Yahoo	25	0.3200		B
Sig.			0.056	0.218

Table 8
Tukey's HSD for precision at DCV 20 P@20 ($\alpha = 0.05$) (human-based evaluation)

Search engines	N	Average precision	Subset for $\alpha = 0.05$	
			1	2
Netscape	25	0.1400	A	
HotBot	25	0.1700	A	B
InfoSeek	25	0.2360	A	B
MSN	25	0.2460	A	B
AlltheWeb	25	0.2480	A	B
Lycos	25	0.2900	A	B
Yahoo	25	0.3280		B
AltaVista	25	0.3340		B
Sig.			0.210	0.124

AWSEEM using the top 100 pages—see Table 10. (Actually AWSEEM 50 only distinguishes Yahoo as the top performer.) The results of Tukey's HSD for human-based evaluation and

Table 9
Tukey's HSD for precision at DCV 20 P@20 ($\alpha = 0.05$) (AWSEEM (50))

Search engines	N	Average precision	Subset for $\alpha = 0.05$	
			1	2
Netscape	25	4.200E−02	A	
HotBot	25	7.800E−02	A	B
InfoSeek	25	9.200E−02	A	B
MSN	25	0.1120	A	B
AlltheWeb	25	0.1120	A	B
Lycos	25	0.1340	A	B
AltaVista	25	0.1380	A	B
Yahoo	25	0.1660		B
Sig.			0.055	0.080

Table 10
Tukey's HSD for precision at DCV 20 P@20 ($\alpha = 0.05$) (AWSEEM (100))

Search engines	N	Average precision	Subset for $\alpha = 0.05$	
			1	2
Netscape	25	8.400E−02	A	
HotBot	25	0.1520	A	B
InfoSeek	25	0.1600	A	B
MSN	25	0.1900	A	B
AlltheWeb	25	0.2020	A	B
Lycos	25	0.2200	A	B
AltaVista	25	0.2300		B
Yahoo	25	0.2680		B
Sig.			0.075	0.211

AWSEEM at DCV 10, 20 using precision (P@10, P@20) reveal that the top two search engines are Yahoo and AltaVista and the worst performing search engine is Netscape.

The same analysis based on recall *at* DCV 10 and 20 (R@10 and R@20) for human-based evaluation forms subsets distinguishing between search engines, whereas analysis of AWSEEM cannot statistically distinguish between these search engines for both top 50 and 100 pages. However, by looking at the rankings the consistency between AWSEEM and human-based evaluation can be observed in determining the top two performers (AltaVista and Yahoo) and the worst two performers (Netscape and HotBot).

In summary, Tukey's HSD shows that human-based and AWSEEM results are in agreement in determining the best and the worst performing search engines. The recommended DCV for Tukey's HSD is 20, since as shown in Fig. 4 at this DCV Pearson r-values saturate and reach their maximum value both for precision and recall performance measures. The statistically significant high Pearson r correlation among precision and recall values of both methods shows that human-based and AWSEEM evaluations can be used interchangeably for measuring the effectiveness of search engines.

In our additional tests, not reported here, we repeated the experiments by limiting the number of terms to 1000 per document (the average document size in the original experiment was 7707 terms). We used two approaches for this: one by only considering the first 1000 words and one by randomly selecting 1000 words of the downloaded pages (if the page size is smaller than 1000 words we used all the words). The Pearson *r* correlation tests show that the results are consistent with that of full text of the Web pages. In other words, the automatic method is applicable with smaller system resources.

6. Conclusions

In this study we present an automatic method for the performance evaluation of Web search engines. We measure the performance of search engines after examining various numbers of top pages returned by the search engines and check the consistency between human and automatic evaluations using these observations. In the experiments we use 25 queries and look at their performance in eight different search engines based on binary relevance judgments of users. Our experiments show a high level of statistically significant consistency between the automatic and human-based assessments both in terms of effectiveness and also in terms of selecting the best and worst performing search engines.

Knowing the most effective Web search engines satisfying the current information-need is important both at a personal and business enterprise level. However, the definition of “the best” changes due to both the changing information-needs of users and the changing quality and nature of search engines. For example, people or business enterprises may work on the solution of different problems at different time instances; and search engines may change their indexing or ranking strategies, and their Web coverage. Accordingly, for different information-needs or with different search technologies the most effective search engine may be different. Hence, search engine performance needs to be tested and this should be done quite often. In this dynamically changing environment our efficient and effective automatic evaluation method, AWSEEM, is a valuable tool, since such evaluations are expensive due to the human time involved.

The method has several practical implications. For example, it can be used in the implementation of

- *Search engine recommenders*: A set of sample queries can be collected from an individual or from users with similar interests and the search engines with the best results can be determined and recommended to the users.
- *Better-quality search engines*: The performance measurement of Web search engines provides motivation to the vendors of these systems to improve their technology.
- *Custom-made meta-search engines*: Best performing search engines can be selected according to the information-needs of a set of users or an individual user and they can be used in the implementation of specialized meta-search engines (Meng, Yu, & Liu, 2002).
- *Benchmarking tools to test search engines in different subject areas*: For subject areas of interest a set of queries can be pooled and the search performance of a particular search engine can be measured. The same approach can also be used to compare the performance of different search engines in these subject areas with respect to each other.

Note that AWSEEM not only measures the effectiveness of search engines, but also effectively defines a meta-search engine (i.e., by ranking and intersecting the downloaded pages with the top pages of search engines). It appears that such a meta-search engine implementation is promising by looking at the consistent AWSEEM and human-based evaluation results.

The human participants of our experiments represent a small portion of the search engine user community at large. A similar argument is also true for the queries used in the experiments. However, since our aim is to show the consistency between human-based and AWSEEM results, this should have no significant effect on the results and similar outcomes are expected with different user groups (Voorhees, 2000). In future experiments, other algorithms may be considered to examine the consistency with human-relevance judgments, since a better ranking procedure may produce more consistent results with user assessments.

One other concern can be the ranking algorithm used in AWSEEM as one might argue that using the same ranking algorithm with a search engine that is being evaluated would cause bias. However, since we collected documents from different search engines and overlap of the documents indexed by each search engine is low (Lawrence & Giles, 1999), the search engine which uses the same algorithm will not retrieve all the documents found relevant by AWSEEM. Furthermore, in AWSEEM for document ranking we use the user information needs (much longer version of user queries), and the term weights used by AWSEEM and a search engine would be different. These factors will lead to different rankings even for the same set of documents. However, it is best not to evaluate a search engine using the same ranking algorithm as AWSEEM.

There are a number of parameters that AWSEEM relies on: n (number of search engines), b (number of pages downloaded from each search engine), s (number of top AWSEEM pages used to measure effectiveness), t or DCV (number of top search engine pages used to measure effectiveness). Our choice of DCV reflects the current Web search engine use practice. In our choice of the number of pages downloaded from each search engine, b, we simulated the approach used in Gordon and Pathak (1999) study. Our current choice of 200 for this parameter appears to be reasonable. Our experiments with the number of top AWSEEM pages used to measure effectiveness show that $s = 100$ is a good choice (for the b, n, and t values used), since it stabilizes the correlation values of human and AWSEEM results. Our choices can be used in other experiments. Other researchers may also perform some sensitivity analysis regarding these parameters before making their final choices. Also in future experiments increasing the number of queries might give more accurate results. AWSEEM provides a generic search engine evaluation approach that can be used in any context. Currently we are using it with TREC data and performing various sensitivity analysis experiments in that environment. The related results will be published in an upcoming article.

Acknowledgements

We would like to thank the Fall 2001 CS533, Information Retrieval Systems, students, and David Davenport of Bilkent University for writing and evaluating the queries used in this study. We would also like to thank our anonymous referees, and Jon M. Patton and Pedrito Uriah Maynard-Zhang for their valuable comments.

References

Abbate, J. (1999). *Inventing the Internet*. Massachusetts: The MIT Press.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: ACM Press.

Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. *Journal of the American Society for Information Science and Technology, 53*(4), 308–319.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Chowdhury, A., & Soboroff, I. (2002). Automatic evaluation of World Wide Web search services. In *Proceedings of the ACM SIGIR conference*, vol. 25 (pp. 421–422).

Gordon, M., & Pathak, P. (1999). Finding information on the World Wide Web: the retrieval effectiveness of search engines. *Information Processing and Management, 35*(2), 141–180.

Hawking, D., Craswel, N., Bailey, P., & Griffiths, K. (2001). Measuring search engine quality. *Information Retrieval, 4*(1), 33–59.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of the ACM SIGIR conference*, vol. 16 (pp. 329–338).

Jansen, B. J., & Pooch, U. (2001). A review of Web searching and a framework for future research. *Journal of the American Society for Information Science and Technology, 52*(3), 235–246.

Kobayashi, M., & Takeda, K. (2000). Information retrieval on the Web. *ACM Computing Surveys, 32*(2), 144–173.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. *Nature, 400*, 107–109.

Meng, W., Yu, C., & Liu, K.-L. (2002). Building efficient and effective metasearch engines. *ACM Computing Surveys, 34*(1), 48–89.

Mowshowitz, A., & Kawaguchi, A. (2002). Assessing bias in search engines. *Information Processing and Management, 35*(2), 141–156.

Salton, G., & Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Soboroff, I., Nicholas, C., & Cahan, P. (2001). Ranking retrieval systems without relevance judgments. In *Proceedings of the ACM SIGIR conference*, vol. 24 (pp. 66–73).

Spink, A., Dietmar, W., Jansen, B. J., & Saracevic, T. (2001). Searching the Web: the public and their queries. *Journal of the American Society for Information Science and Technology, 52*(3), 226–234.

Spink, A., Jansen, B. J., Wolfram, D., & Saracevic, T. (2002). From E-sex to E-commerce: Web search changes. *Computer, 35*(3), 107–109.

Voorhees, E. M. (2000). Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing and Management, 36*(5), 697–716.